Going from -omics outputs to dynamic networks

David Shorthouse – d.shorthouse@ucl.ac.uk

This is an example workflow that can take you from a "list of genes" or equivalent, to a network that might let you start understanding how the dynamics of the system behave.

Note: I am not necessarily using best practises for enrichment analysis etc. but am doing something "quick and dirty" to show you how to work with a system. Ideally you would script up eg. extraction of your differential genes rather than using excel, for reproducibility reasons.

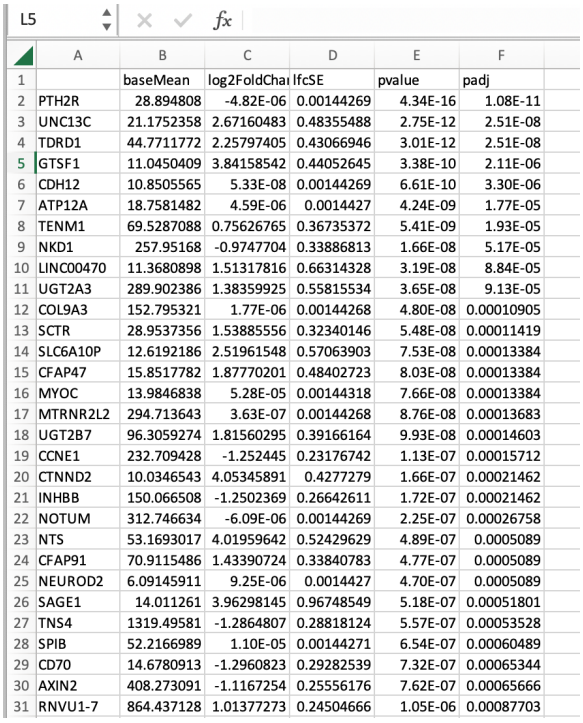
The workflow can be broadly described through these steps (the first two of which are part of a standard informatics workflow):

1. Collect your list of genes or biologically relevant targets
2. Perform an enrichment or equivalent to identify pathways or broader biology
3. Visualise and analyse the connectivity of your targets
4. Selecting and building a model of the system
5. Analysing model dynamics – This is covered in the rest of the course.

In this example I am taking the results of a differential expression analysis performed by DESeq2 on a set of patients mutant or wild-type for a gene (KCNQ3), that we know drives metastasis in oesophageal adenocarcinoma.

1. Perform your differential expression analysis/mutational analysis etc:

Here we open the DESeq2 results as an excel spreadsheet, sort by adjusted pvalue, and extract genes with a padj <= 0.05.



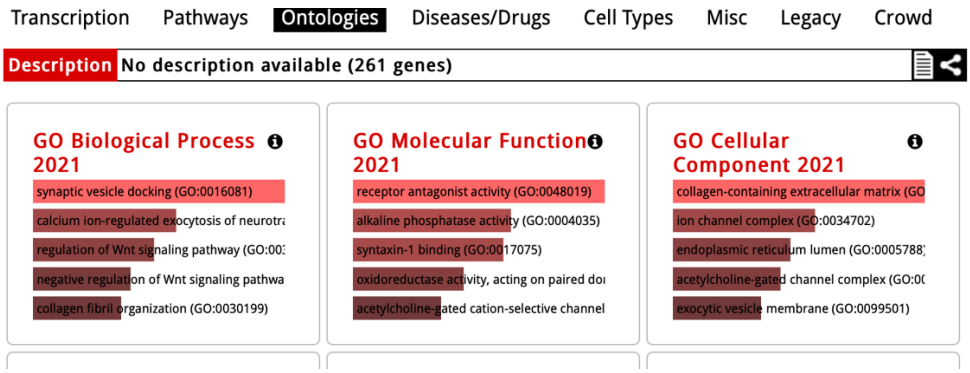
L5

	A	B	C	D	E	F
1		baseMean	log2FoldChai	lfcSE	pvalue	padj
2	PTH2R	28.894808	-4.82E-06	0.00144269	4.34E-16	1.08E-11
3	UNC13C	21.1752358	2.67160483	0.48355488	2.75E-12	2.51E-08
4	TDRD1	44.7711772	2.25797405	0.43066946	3.01E-12	2.51E-08
5	GTSF1	11.0450409	3.84158542	0.44052645	3.38E-10	2.11E-06
6	CDH12	10.8505565	5.33E-08	0.00144269	6.61E-10	3.30E-06
7	ATP12A	18.7581482	4.59E-06	0.0014427	4.24E-09	1.77E-05
8	TENM1	69.5287088	0.75626765	0.36735372	5.41E-09	1.93E-05
9	NKD1	257.95168	-0.9747704	0.33886813	1.66E-08	5.17E-05
10	LINC00470	11.3680898	1.51317816	0.66314328	3.19E-08	8.84E-05
11	UGT2A3	289.902386	1.38359925	0.55815534	3.65E-08	9.13E-05
12	COL9A3	152.795321	1.77E-06	0.00144268	4.80E-08	0.00010905
13	SCTR	28.9537356	1.53885556	0.32340146	5.48E-08	0.00011419
14	SLC6A10P	12.6192186	2.51961548	0.57063903	7.53E-08	0.00013384
15	CFAP47	15.8517782	1.87770201	0.48402723	8.03E-08	0.00013384
16	MYOC	13.9846838	5.28E-05	0.00144318	7.66E-08	0.00013384
17	MTRNR2L2	294.713643	3.63E-07	0.00144268	8.76E-08	0.00013683
18	UGT2B7	96.3059274	1.81560295	0.39166164	9.93E-08	0.00014603
19	CCNE1	232.709428	-1.252445	0.23176742	1.13E-07	0.00015712
20	CTNND2	10.0346543	4.05345891	0.4277279	1.66E-07	0.00021462
21	INHBB	150.066508	-1.2502369	0.26642611	1.72E-07	0.00021462
22	NOTUM	312.746634	-6.09E-06	0.00144269	2.25E-07	0.00026758
23	NTS	53.1693017	4.01959642	0.52429629	4.89E-07	0.0005089
24	CFAP91	70.9115486	1.43390724	0.33840783	4.77E-07	0.0005089
25	NEUROD2	6.09145911	9.25E-06	0.0014427	4.70E-07	0.0005089
26	SAGE1	14.011261	3.96298145	0.96748549	5.18E-07	0.00051801
27	TNS4	1319.49581	-1.2864807	0.28818124	5.57E-07	0.00053528
28	SPIB	52.2166989	1.10E-05	0.00144271	6.54E-07	0.00060489
29	CD70	14.6780913	-1.2960823	0.29282539	7.32E-07	0.00065344
30	AXIN2	408.273091	-1.1167254	0.25556176	7.62E-07	0.00065666
31	RNVU1-7	864.437128	1.01377273	0.24504666	1.05E-06	0.00087703

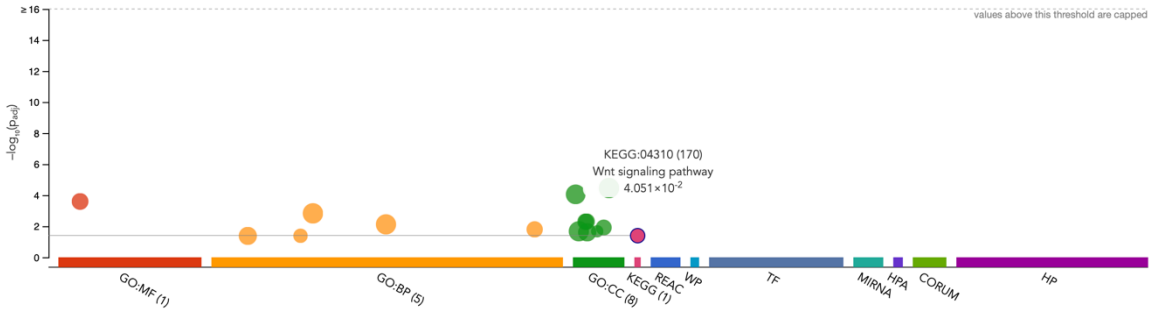
This gives me a list of ~250 genes. If you have many genes, you may need to perform additional steps like using log2(fold change) cut-offs to get the list down. A general rule for enrichment analysis is that a couple of hundred genes or less is optimal.

2. Identify some biology associated with your genes of interest:

Put this gene list into eg. enrichr (https://maayanlab.cloud/Enrichr/), gprofiler (https://biit.cs.ut.ee/gprofiler/gost), GSEA or equivalent. This will give you a feel for which pathways are altered in your conditions. For enrichr I generally look at GO pathways, reactome, and KEGG:



Gprofiler gives you a nice interactive interface to see which terms are enriched in your gene list:



Both of these tools allow you to download results as spreadsheets that you can plot separately, or can be run on the command line to allow you to perform the plotting yourself.

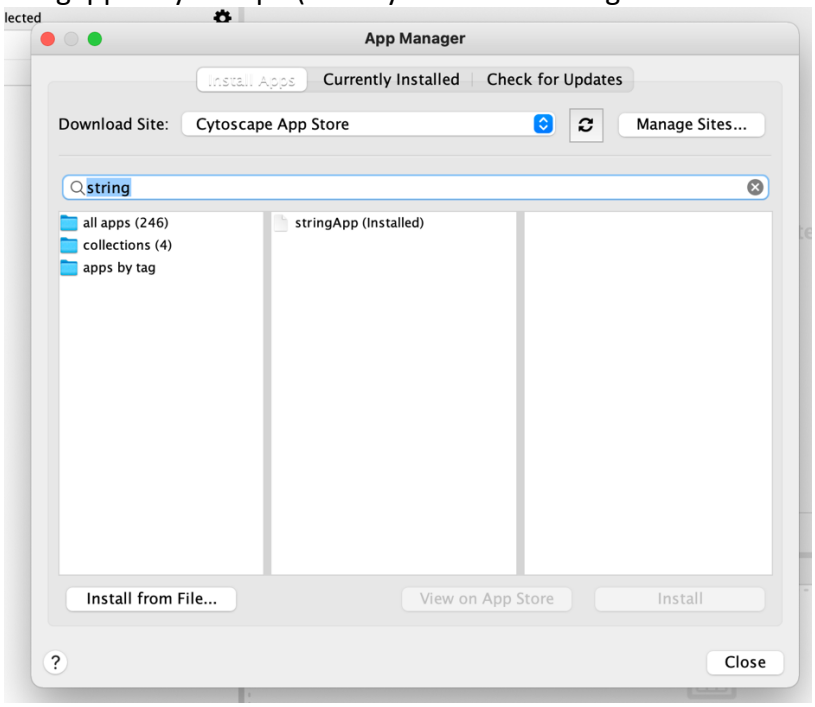
GSEA and enrichr are included in the GSEApy python package: https://gseapy.readthedocs.io/en/latest/introduction.html

Gprofiler can be run using the gprofiler-official python package: https://pypi.org/project/gprofiler-official/

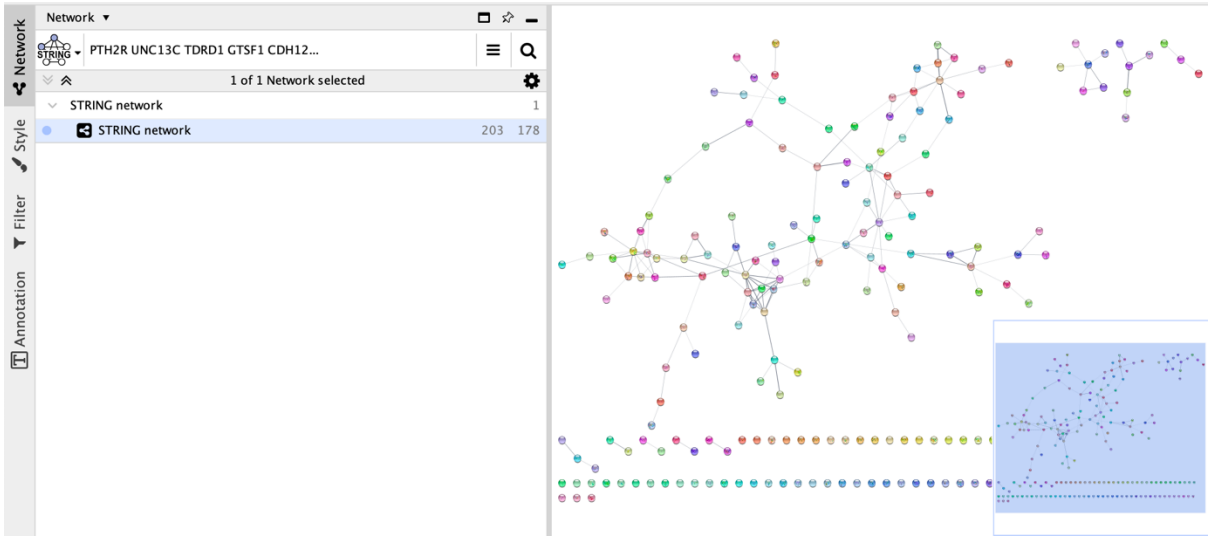
This analysis can give you a feel for what is changed in your conditions, can help you subset your genes down, and can provide you with a lot of data that you can download and replot. In this instance we identify that in our mutant patients, there is an enrichment for Wnt signalling, epithelial to mesenchymal transition pathways, and extracellular matrix dynamics.

3. Perform a simple visualisation/connectivity analysis using cytoscape:

Download the stringapp in cytoscape (Note: you can also use genemania for this):



Stringapp includes a functionality to insert a list of genes and get a connectivity graph for them. In this case, if I paste my gene list into this window, String goes off and "joins up" these genes where it can, you can tweak parameters such as the inclusion of "linker" nodes that are not in your query:

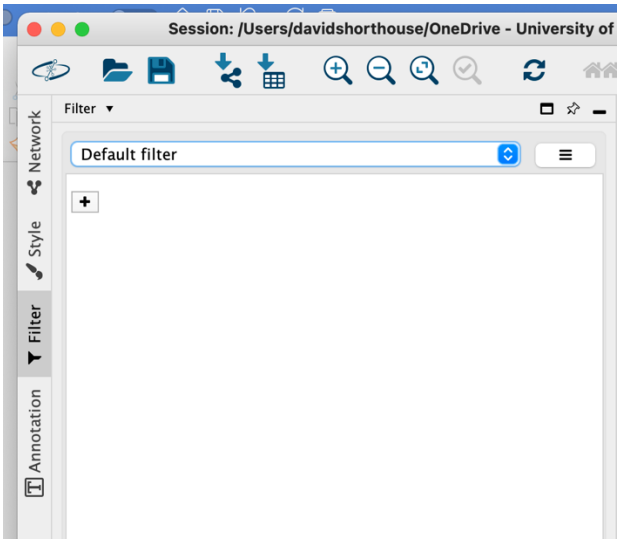


You can now explore these clusters using cytoscape, overlay expression data, colour by function, or subset the network. (A good tutorial for working with cytoscape is here: https://cytoscape.org/cytoscape-tutorials/protocols/functional-enrichment)

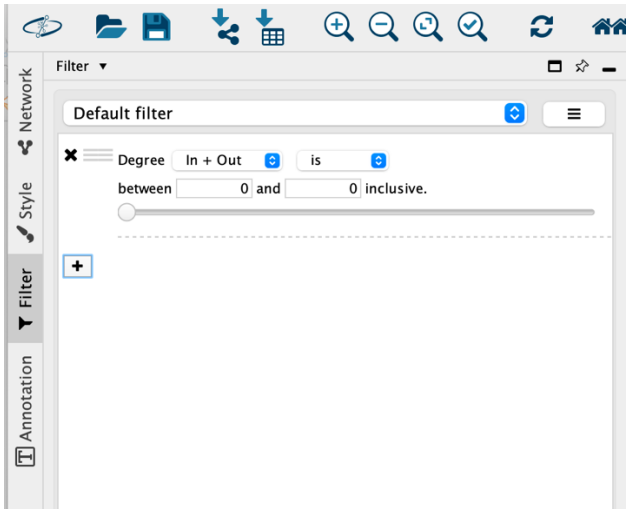
Some basics in cytoscape though are to:

A) Prune the network by removing all nodes that are unconnected.

To do this go to the “filter” tab:



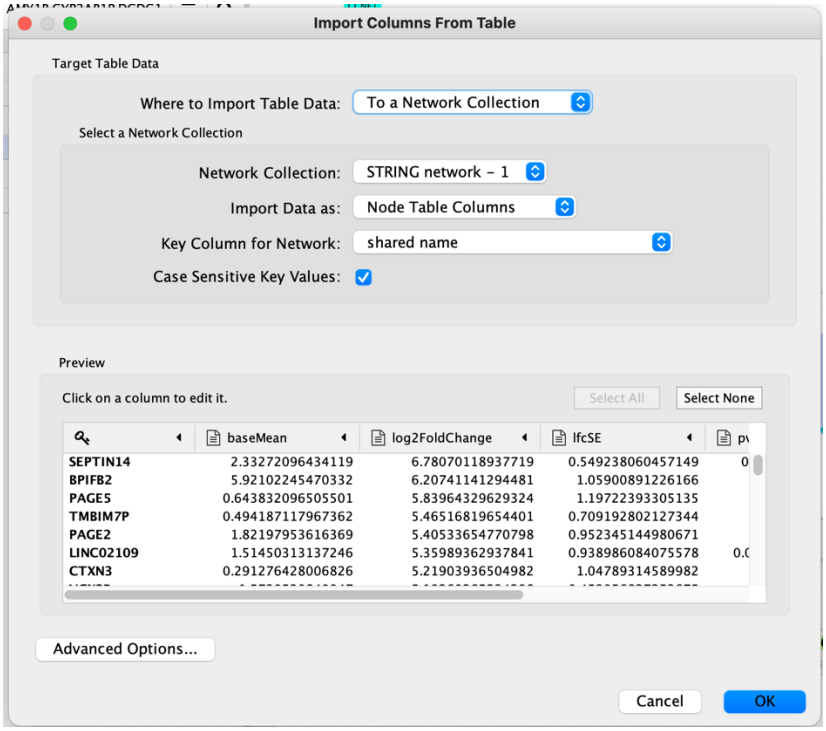
Create a new degree filter:



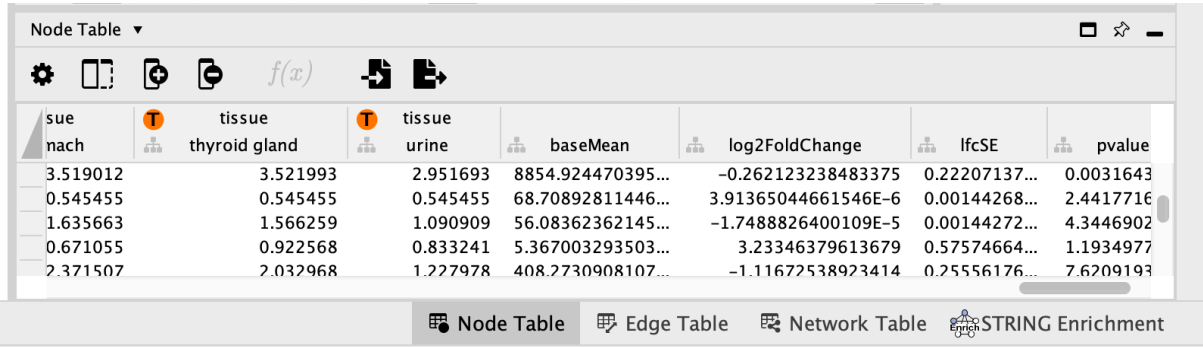
You can now adjust these parameters to choose nodes with a certain degree of connectivity. You can then hide, delete, or colour these nodes differently after selection.

B) Import expression data to overlay.

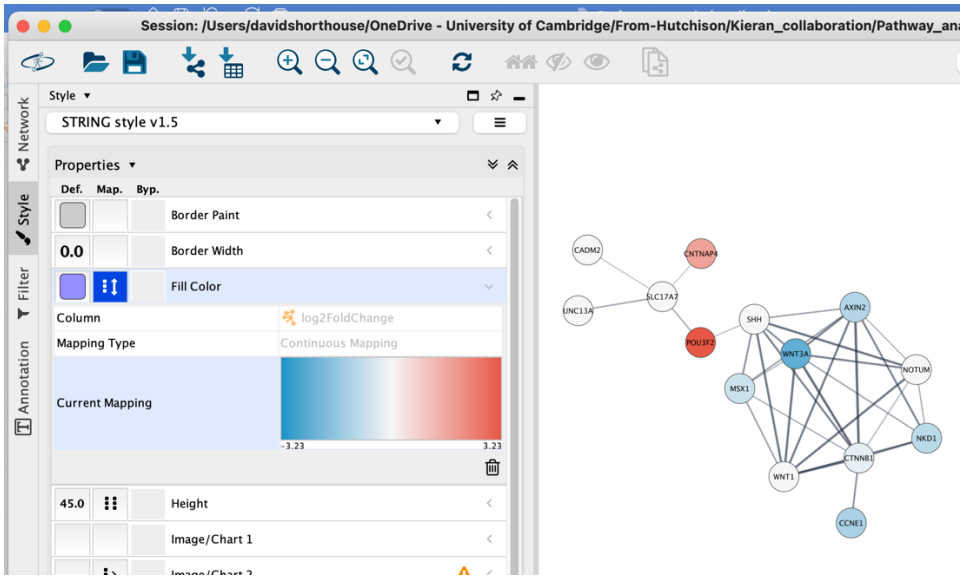
We can overlay fold change, or any other data on top of our network. To do this, go to Import -> Table from file



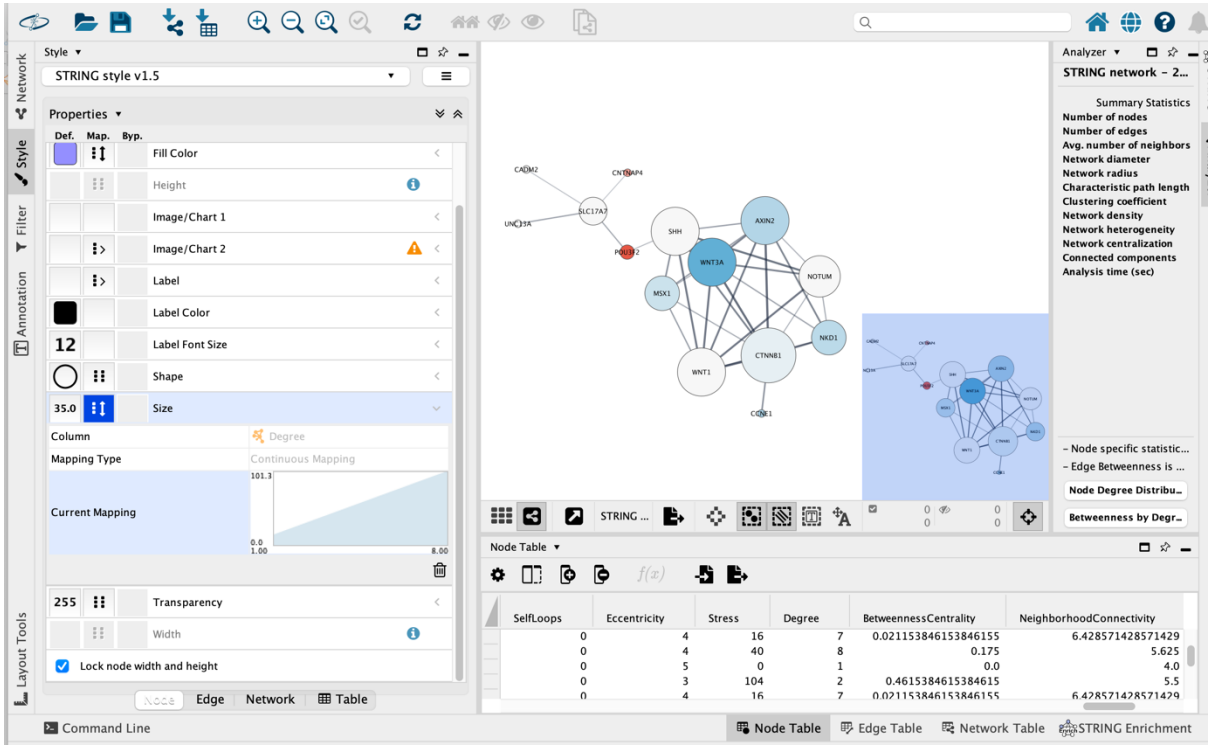
Choose a network to import to, import the data as node table columns, you need to identify which node property is matched to your spreadsheet (in this case its “display name”), and probably untick “Case Sensitive Key Values”. This should have imported the data from your differential expression on top of the network, with each node having the values from the table. The Node Table should look similar to the below:



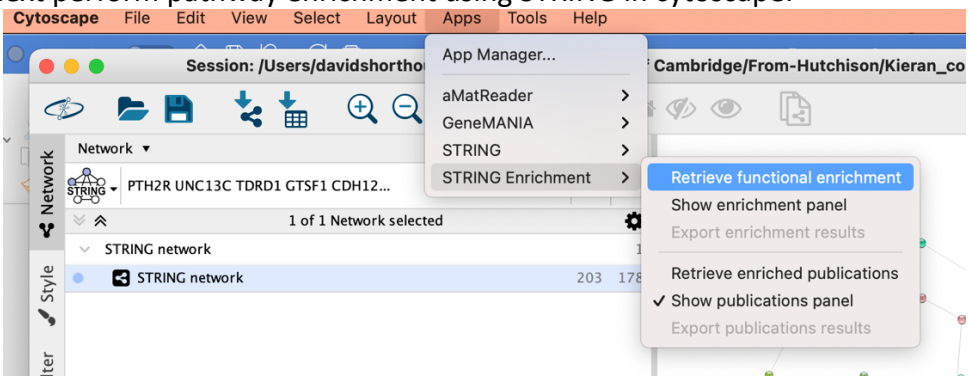
We can now do things like colour our nodes by fold change. Go to the style tab on the left, choose or create a new style, go to fill colour and click the “Mapping” button – this will then allow you to choose a continuous mapping on a column of the node table



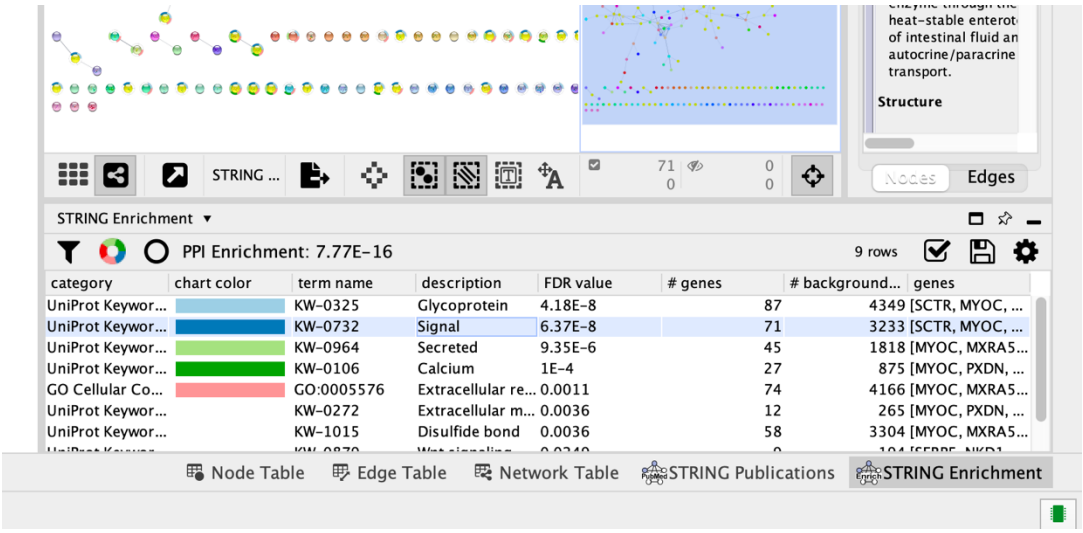
We can also include network properties in our styles. First we have to calculate our network properties by using Tools -> Analyze Network. This will add some columns to your node table, including one called “degree” which is the number of incoming and outgoing connections. We can now change the size of the nodes by this parameter by following the same steps as for fill colour:



We can next perform pathway enrichment using STRING in cytoscape:

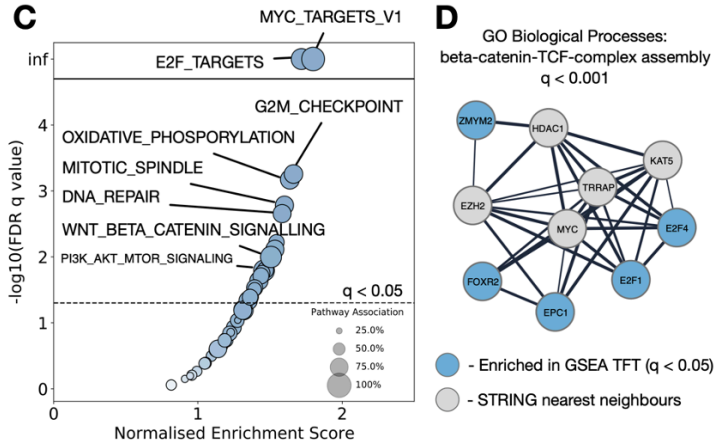


The STRING enrichment tool lets you calculate enriched pathways in networks. You can calculate for all of the genes (which should give you same results as if you put the gene list into gprolifer etc.), or you can select subnetworks to see what they do. Also – if you have removed unconnected nodes, this will change your enrichment results. An advantage here is that you can select enriched terms and then highlight/select the nodes that are associated with the pathway. These nodes can then be coloured, reclustered, downloaded as SBML, publications can be looked at etc.



category	chart color	term name	description	FDR value	# genes	# background...	genes
UniProt Keywor...		KW-0325	Glycoprotein	4.18E-8	87	4349	[SCTR, MYOC, ...
UniProt Keywor...		KW-0732	Signal	6.37E-8	71	3233	[SCTR, MYOC, ...
UniProt Keywor...		KW-0964	Secreted	9.35E-6	45	1818	[MYOC, MXRA5...
UniProt Keywor...		KW-0106	Calcium	1E-4	27	875	[MYOC, PXDN, ...
GO Cellular Co...		GO:0005576	Extracellular re...	0.0011	74	4166	[MYOC, MXRA5...
UniProt Keywor...		KW-0272	Extracellular m...	0.0036	12	265	[MYOC, PXDN, ...
UniProt Keywor...		KW-1015	Disulfide bond	0.0036	58	3304	[MYOC, MXRA5...

You can now colour nodes by enriched terms, subset nodes, get publications etc. and generally this can be used to direct the building of your first model. In many cases, I just plot the outputs of these and include them in my publications. For example, this network of differentially expressed genes that are involved in beta-catenin signalling in these patients:



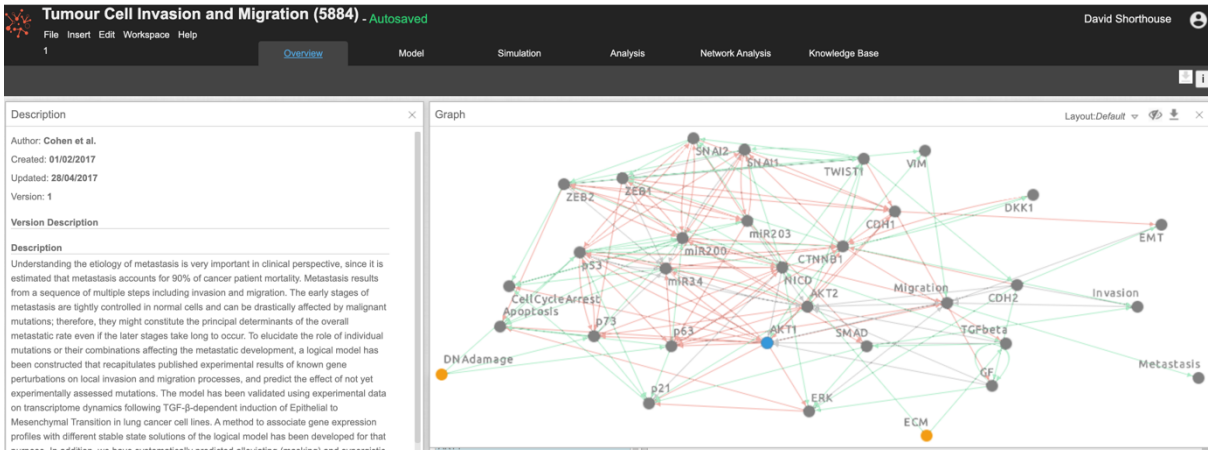
4. Selecting and building a model of the system:

There are 2 ways we can start to do some modelling:

a) We take an existing model, adapt it based on our map/data, and study the dynamics.
b) We build a new model from scratch, using our map from cytoscape as a basis.

In our existing example of mutant KCNQ3 in metastasis of oesophageal adenocarcinoma:

Now that we know we're interested in beta-catenin signalling (the Wnt pathway), we can see whether someone has already generated a model we can start from – in this case, putting "WNT" into the search in cellcollective (cellcollective.org) finds us a model of tumor invasion and metastasis that is heavily focussed on the Wnt pathway, many of the nodes overlap with those in our network, so this may be a great starting point:



We can now use cellcollective to analyse the model, add new nodes, adjust nodes to meet our criterion, study stability, etc. We could also look to download a map from The Atlas of Cancer Signalling and start from there, or go a number of different directions.

If we want to build a new model, there are many ways to do this, but a starting point is to find a review or big signalling paper that discusses the pathway of interest, and start from there, using your cytoscape map as a guide – eg this model of the Wnt pathway built in the biomodelanalyzer based on this paper https://doi.org/10.1016/j.cell.2018.03.035: